Logistic Factor Analysis Vignette

Wei Hao, Minsun Song, John D. Storey April 15, 2025

1 Introduction

Logistic Factor Analysis (LFA) [1]. Briefly, LFA fits a latent variable model on categorical (i.e. SNP genotypes coded as 0, 1, and 2) data by modeling the logit transformed binomial parameters in terms of latent variables. The resulting "logistic factors" are analogous to principal components, but fit into a convenient likelihood based model. As a result, the logistic factors can power a number of other analyses.

2 Sample usage

We include a sample real dataset with the package as the variable hgdp_subset—a small subset of the HGDP genotypes. The row names are the rsids for the SNPs and the column names are coarse geographical labels for the individuals.

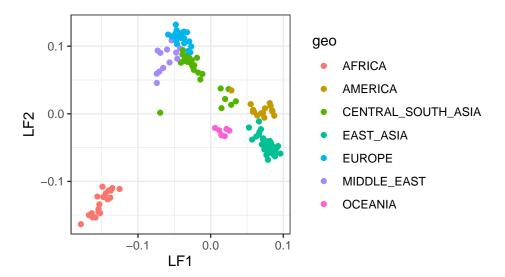
```
library(lfa)
dim(hgdp_subset)
## [1] 5000 159
```

2.1 lfa

The lfa function has two required arguments. The first is the genotype matrix, and the second is the number of logistic factors including the intercept.

```
LF <- lfa(hgdp_subset, 4)
dim(LF)
## [1] 159
head(LF)
##
               [,1]
                           [,2]
                                        [,3] [,4]
## [1,] -0.03635636 0.10431465 -0.026157330
## [2,] -0.04618695 0.10283828 -0.014107285
## [3,] -0.02800495 0.09071295 -0.030261474
## [4,] -0.03513497 0.08332765 -0.002122198
                                                1
## [5,] -0.03064114 0.07938743 -0.021696325
                                                1
## [6,] -0.02442314 0.06953981 -0.005312549
```

We can plot the first two logistic factors and color by geographical information:



One aspect of lfa is that the return value is a matrix of logistic factors, thus, an important part of subsequent analysis is to keep your matrix of logistic factors to pass as an argument.

2.2 af

Given a genotype matrix and logistic factors, the af function computes the individual-specific allele frequencies

Since the calculation is independent at each locus, you can pass a subset of the genotype matrix as an argument if you aren't interested in all the SNPs.

Given the allele frequencies, you can do some other interesting calculations—for example, compute the log-likelihood for each SNP.

3 Data Input

The best way to load genotypes is by using the function read_plink, from the genio package, which assumes that you have binary PLINK formatted genotypes. The binary PLINK format uses files: a .bed for the genotypes, a .bim for the genotype information, and a .fam for the individuals information. read_plink takes as an argument the prefix for your three files. A BEDMatrix object (from the eponymous function and package) is also supported, and can result in reduced memory usage (at a small runtime penalty).

References

[1] Wei Hao, Minsun Song, and John D. Storey. Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, 32(5):713–721, March 2016.